What is the Area under the Receiver Operating Curve (AUC) in Prediction?

Overview:

When a predictive model or test is developed, we often see an **AUC** or **c-statistic** (these two terms may be used interchangeably) reported to describe its discriminatory ability. AUC always has a value between 0.50 and 1.0. *What exactly is this quantity and how do we interpret it?*

Motivation:

Suppose we want to predict whether a patient has coronary heart disease (CHD). We have used the data from a set of patients from a specific center and trained a model (such as a **logistic regression** or **Random Forest**....) that generated a CHD prediction for each patient using information such as demographics and lab values, such as LDL levels.

Here is a snippet of the data set with true and predicted CHD status:

Patient	**age**	**LDL**	**CHD status**	**Prediction**
1	65	144	1	0.31
2	47	103	0	0.10
3	71	120	0	0.28
4	62	176	1	0.55
5	60	190	1	0.46
6	59	148	0	0.01

Of course, we want to assess how well this model actually predicts CHD. To do this, we can compare the prediction generated by the model with the true CHD status observed for each patient. If we do this comparison in the same sample used to build the model, this is called *in-sample* performance. This is the weakest form of assessing model performance because we have used the data to both build the model and assess discrimination, which tends to greatly inflate estimates of performance. If we use a completely new cohort that was not involved at all in model building, this is called *external validation*. "External" here means external to the model-building process and could mean that you are validating the model using: 1) a portion of the existing data set not used in model training (i.e.: a holdout test set), 2) a data set collected from the same clinic at a different point in time, 3) a data set from a different clinic or institution, or 4) a data set from a different population entirely.

How do we use these predictions? At every possible threshold of probability, we can create a 2x2 table of predicted vs. true CHD status. For example, if I set a threshold of 0.20 where a patient with a prediction > 0.20 was predicted to have CHD, we would have the following:

	True Status	
	CHD	**No CHD**
CHD predicted	123	48
No CHD predicted	72	805

Similarly, if I set the threshold to be 0.50 such that patients with predictions > 0.50 would be predicted to have CHD, we would have the following 2x2 table:

	True Status	
	CHD	**No CHD**
CHD predicted	83	35
No CHD predicted	112	818

We can generate a 2x2 table for every possible threshold from 0 to 1.0. Of course, as the probability threshold for CHD increases, fewer and fewer patients will be predicted to have CHD. For each 2x2 table generated, we can compute the sensitivity and specificity corresponding that a particular threshold. Recall that sensitivity is the proportion of patients with CHD who have CHD predicted, and specificity is the proportion of patients without CHD who do not have CHD predicted.

	True Status	
	CHD	**No CHD**
CHD predicted	83	35
No CHD predicted	112	818

Sensitivity=83/(83+112) = 0.43

Specificity=818/(818+35) = 0.96

Now that we have sensitivity and specificity for all possible thresholds, we want to summarize this succinctly to describe how well the model discriminates CHD from non-CHD cases. This is where the ROC curve and the area under the ROC curve (AUC) are useful.

- **The ROC curve:** The ROC curve is a plot of sensitivity (Y-axis) vs. 1-specificity (X-axis) across all possible thresholds. This curve allows you to see the tradeoff between sensitivity and specificity. **Figure 1** provides an example of an ROC curve:

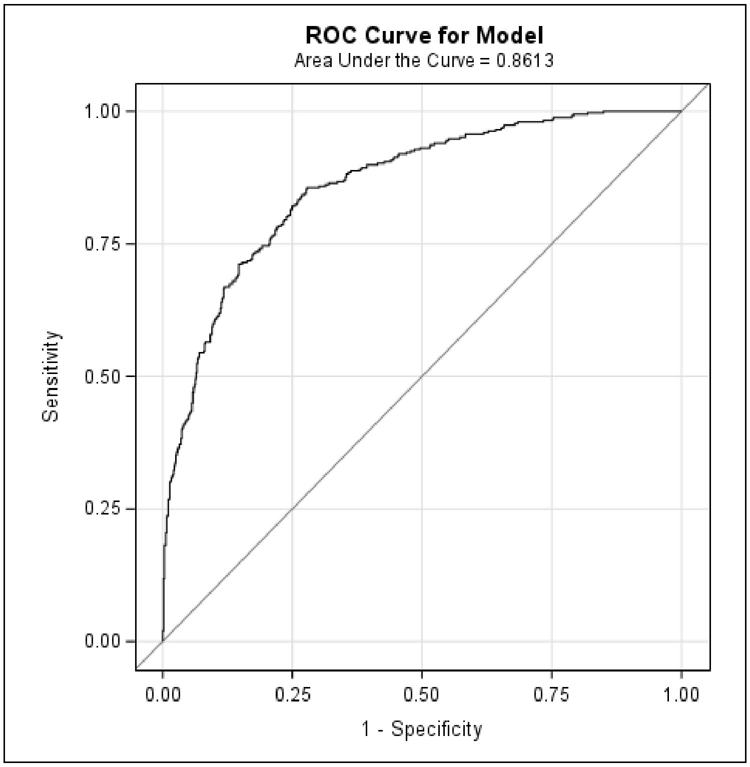

Figure 1: ROC curve

At each threshold, we look at the proportion of CHD with predicted value=1 (sensitivity) vs the proportion of non-CHD with prediction=1 (1-specificity). At every threshold, we want the proportion of predicted values =1 to be higher for patients with CHD vs. non-

CHD. If the sensitivity is close to 1 and 1-specificity close to zero at any thresholds, then at least one point in the ROC curve will be high in the upper left corner at (0,1) and yield a very high AUC value.

- The **AUC** is the computed area under the ROC curve. If we achieve perfect discrimination with our predictions, we will have a computed AUC value of 1.0. If our model predicts no better than rolling a die to decide who has CHD, the computed AUC will be 0.50.

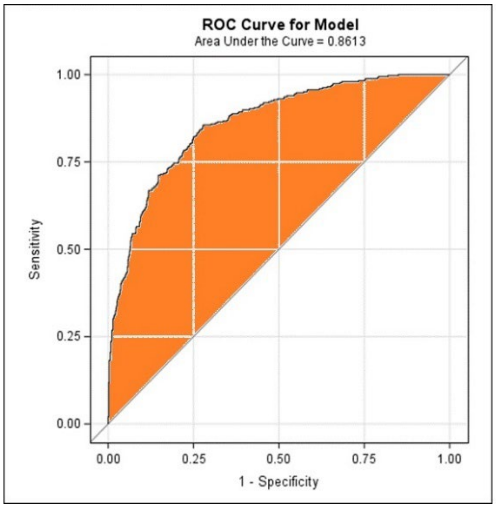

Figure 2: ROC curve with the area (AUC) shaded in orange.

How do we interpret this AUC? Consider every possible pairing of a patient with CHD to a patient without CHD. If we compute the proportion of instances where the patient with CHD has a higher predicted probability for CHD compared to the patient without CHD in that pair, we arrive at the AUC value.

In our example, in a randomly selected (CHD, no CHD) pair, the patient with CHD will have a higher predicted probability than the patient without CHD 86% of the time.

What is a “good” AUC value?

AUC 0.90-1.00	Excellent discrimination
AUC 0.80-0.90	Good
AUC 0.70-0.80	Moderate
AUC 0.60-0.70	Poor
AUC 0.50-0.60	No discrimination

In our example, the computed AUC is equal to 0.86, which is considered good discrimination performance.

Where can I read more about this?

There are many **excellent online resources** that will allow you to more fully understand.